Claims

What is claimed is:

1. A method for pattern discovery on an input sequence comprising a plurality of elements, the method comprising the steps of:

determining a plurality of first motifs from the input sequence, each first motif comprising at least one element from the input sequence;

concatenating each of the plurality of first motifs with another of the plurality of first motifs to create a plurality of concatenated motifs;

removing selected motifs of the concatenated motifs and the first motifs.

2. The method of claim 1, further comprising the step of selecting motifs of the concatenated motifs and the first motif for removal based on at least one predetermined criteria.

3. The method of claim 1, wherein the step of removing comprises removing suffix motifs in the concatenated motifs and the first motifs.

4. The method of claim 3, wherein each motif in the concatenated motifs and the first motifs has an associated location list, and wherein the step of removing suffix motifs comprises the steps of:

offsetting each location list for each of the motifs in the concatenated motifs and the first motifs to zero;

checking each location list for each of the motifs in the concatenated motifs and the first motifs to determine location lists that are the same; and

augmenting a motifs that have the same location list to create at least one new motif.

5. The method of claim 1, wherein the step of removing comprises removing redundant motifs in the concatenated motifs and the first motifs.

6. The method of claim 5, wherein each motif in the concatenated motifs and the first motifs has an associated location list, and wherein the step of removing redundant motifs comprises the steps of:

determining any motif whose location list is a union of other location lists associated with motifs in the concatenated motifs and the first motifs; and

removing any motif whose location list is a union of other location lists associated with motifs in the concatenated motifs and the first motifs.

7. The method of claim 1, wherein the step of removing comprises removing selected motifs in the concatenated motifs and the first motifs if the selected motifs do not occur in the concatenated motifs and the first motifs more than a predetermined number of times.

8. The method of claim 1, further comprising the step of:

performing the steps of concatenating and removing until no new motifs are generated.

9. The method of claim 1, wherein:

each first motif is a solid element motif;

the step of determining a plurality of first motifs comprises the steps of:

determining a plurality of solid element motifs, each solid element motif comprising at least one element from the input sequence; and

creating a plurality of second motifs by adding at least one don't care element to each of the solid element motifs;

the step of concatenating further comprises the steps of:

selecting a motif from the solid element and second motifs;

concatenating the selected motif with another selected motif from the solid element and second motifs; and

performing the process of selecting and concatenating until each motif has been concatenated with another motif;

the method further comprises the steps of:

trimming the solid element, second, and concatenated motifs; and

performing the steps of concatenating and trimming until no new motifs are generated.

10. The method of claim 9, further comprising the step of creating flexible motifs from the first motifs.

11. The method of claim 1, further comprising the step of creating flexible motifs from the first motifs.

12. The method of claim 1, wherein each element of the input sequence comprises a character from an alphabet.

13. The method of claim 1, wherein at least one element of the input sequence comprises a set of characters.

14. The method of claim 1, wherein each element of the input sequence comprises a real number.

15. The method of claim 8, wherein remaining motifs of the concatenated motifs and the first motifs form a basis set of motifs and wherein the method further comprises the steps of:

determining a plurality of motif sets from a plurality of selected motifs, the selected motifs selected from a plurality of basis motifs, wherein the plurality of selected motifs all begin with a selected element;

determining unique intersection sets from the plurality of motif sets; and

determining redundant motifs from the intersection sets and the motif sets.

16. A computer system comprising:

a memory that stores computer-readable code;

a processor operatively coupled to the memory, the processor configured to implement the computer-readable code, the computer-readable code configured to:

determine a plurality of first motifs from the input sequence, each first motif comprising at least one element from the input sequence;

concatenate each of the plurality of first motifs with another of the plurality of first motifs to create a plurality of concatenated motifs; and

remove selected motifs of the concatenated motifs and the first motifs.

17. An article of manufacture comprising:

a computer readable medium having computer-readable code means embodied thereon, the computer-readable program code means comprising:

a step to determine a plurality of first motifs from the input sequence, each first motif comprising at least one element from the input sequence;

a step to concatenate each of the plurality of first motifs with another of the plurality of first motifs to create a plurality of concatenated motifs; and

a step to remove selected motifs from the concatenated motifs and the first motifs.

18. A method, comprising:

determining a plurality of motif sets from a plurality of selected motifs, the selected motifs selected from a plurality of maximal irredundant motifs, wherein each of the plurality of selected motifs begins with a selected element;

determining intersection sets from the motif sets; and

determining redundant motifs from the intersection sets and the motif sets.

19. The method of claim 18, wherein the step of determining intersection sets from the motif set further comprises the step of determining unique intersection sets from the motif set.

20. The method of claim 19, further comprising the step of performing the steps of determining a plurality of motif sets, determining unique intersection sets, and determining redundant motifs for a plurality of subsets of the maximal irredundant motifs.

21. The method of claim 19, wherein the unique intersection sets are determined through a solution to a Set Intersection Problem (SIP).

22. The method of claim 19, wherein step of determining the unique intersection sets further comprises the steps of:

selecting one motif from the selected maximal irredundant motifs;

determining an intersection set of motif sets that contain the selected motif;

determining if the intersection set has not been previously determined;

naming the intersection set as a unique intersection set when the intersection set has not been previously determined; and

performing the steps of selecting one motif, determining an intersection set, determining if the intersection set has not been previously determined, and naming the intersection set until there are no more motifs from the selected maximal irredundant motifs.

23. The method of claim 19, wherein each motif has an associated location list, and wherein the step of determining redundant motifs from the intersection sets and the motif sets further comprises the steps of:

for each intersection set, performing an intersection of the motifs in the intersection set and performing an intersection of the location lists for the motifs in the intersection set; and

for each motif set, performing an intersection of the sets in the motif set to determine which motifs are common to the sets, performing an intersection of the common motifs and performing an intersection of the location lists for the common motifs.

24. A computer system comprising:

a memory that stores computer-readable code;

a processor operatively coupled to the memory, the processor configured to implement the computer-readable code, the computer-readable code configured to:

determine a plurality of motif sets from a plurality of selected motifs, the selected motifs selected from a plurality of maximal irredundant motifs, wherein each of the plurality of selected motifs begins with a selected element;

determine intersection sets from the motif sets; and

determine redundant motifs from the intersection sets and the motif sets.

25. An article of manufacture comprising:

a computer readable medium having computer-readable code means embodied thereon, the computer-readable program code means comprising:

a step to determine a plurality of motif sets from a plurality of selected motifs, the selected motifs selected from a plurality of maximal irredundant motifs, wherein each of the plurality of selected motifs begins with a selected element;

a step to determine intersection sets from the motif sets; and

a step to determine redundant motifs from the intersection sets and the motif sets.

26. A method for pattern discovery on an input sequence comprising a plurality of elements, the method comprising the steps of:

determining a plurality of basis motifs from the input sequence, each of the basis motifs being a maximal irredundant motif; and

determining a plurality of maximal redundant motifs from the plurality of basis motifs.